AFHRL-TP-90-61

(2)

AIR FORCE

HUMAN RESOURCES

INTERVIEW TESTING AS A WORK SAMPLE MEASURE OF JOB PROFICIENCY

Jerry W. Hedge

Personnel Decisions Research Institutes, Incorporated
43 Main Street SE, Suite 405
Minneapolis, Minnesota 55414

Mark S. Teachout

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Frances J. Laue

Universal Energy Systems, Incorporated
8961 Tesoro Drive, Suite 600
San Antonio, Texas 78217

November 1990

Interim Technical Paper for Period January 1989 – July 1990

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Colonel, USAF
Chief, Training Systems Division

This technical paper is printed as received and has not been edited by the AFHRL Technical Editing staff. The opinions expressed herein represent those of the author and not necessarily those of the United States Air Force.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1990	Interim Paper - January 1989 to July 1990

4. TITLE AND SUBTITLE

Interview Testing as a Work Sample Measure of Job Proficiency

5. FUNDING NUMBERS

C - F41689-86-D-0052
PE - 63227F
PR - 2922
TA - 01
WU - 01

6. AUTHOR(S)

Jerry W. Hedge
Mark S. Teachout
Frances J. Laue

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Universal Energy Systems, Incorporated
8961 Tesoro Drive, Suite 600
San Antonio, Texas 78217

8. PERFORMING ORGANIZATION REPORT NUMBER

AFHRL-TP-90-61

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

Training Systems Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The development of criteria to measure individual proficiency is a necessary prerequisite for most personnel decisions. As part of the Air Force's Job Performance Measurement project, performance measures were developed and administered to more than 1400 enlisted airmen across eight Air Force specialties (AFSs). Included in these measures were two work sample tests (i.e., hands-on, interview), four rating forms of varying specificity, and job knowledge tests. The current research effort centers on Interview Testing as an alternative to the more costly and time-consuming Hands-on Testing. Data analyses revealed that correlations between the two work sample tests ranged from moderate (.46) to high (.84) across the eight AFSs. Also, the patterns of relationships between each work sample test and the series of related measures were quite similar. These results suggest that the interview approach offers some promise as a work sample measure of job proficiency. The discussion addresses the usefulness of this approach for performance measurement, validation of selection/classification procedures, evaluating training programs, and identifying training deficiencies.

14. SUBJECT TERMS

criterion-development
hands-on testing
interview
job performance
job proficiency
performance measurement
surrogate measure
work sample

15. NUMBER OF PAGES

66

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 90-61

November 1990

INTERVIEW TESTING AS A WORK SAMPLE MEASURE OF JOB PROFICIENCY

Jerry W. Hedge

Personnel Decisions Research Institutes, Incorporated
43 Main Street SE, Suite 405
Minneapolis, Minnesota 55414

Mark S. Teachout

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Frances J. Laue

Universal Energy Systems, Incorporated
8961 Tesoro Drive, Suite 600
San Antonio, Texas 78217

Reviewed and submitted for publication by

James B. Bushman, Major, USAF
Chief, Training Assessment Branch
Training Systems Division

SUMMARY

The Air Force Human Resources Laboratory has developed a performance measurement technology to evaluate selection/classification methodologies, training programs, and research and development (R&D) efforts. This Job Performance Measurement System (JPMS) includes work sample tests, rating forms, questionnaires, and job knowledge tests. The work sample approach consists of both a Hands-on Test and an Interview Test.

JPMS data have been collected for eight Air Force specialties (AFSs). This paper presents the results of test-level and task-level analyses on the viability of the interview as a work sample measurement methodology. Correlational relationships between the two approaches were moderate to high. However, significant differences were found between hands-on and interview mean values. Implications of these findings and suggestions for future R&D efforts are discussed.

PREFACE

For the last several years, the Air Force Human Resources Laboratory (AFHRL) has been participating in a Joint-Service Job Performance Measurement Project. As part of this project, the AFHRL has developed an Interview Testing work sample approach to performance measurement. This paper provides results of analyses on the viability of the interview as a measurement methodology. This work was conducted under Contract No. F41689-86-D-0052, awarded to Universal Energy Systems, Inc.

TABLE OF CONTENTS

TABLE OF CONTENTS (Concluded)

LIST OF TABLES

LIST OF TABLES (Concluded)

LIST OF FIGURES

INTERVIEW TESTING AS A WORK SAMPLE MEASURE OF JOB PROFICIENCY

I. INTRODUCTION

Accurate selection, classification, training, and performance evaluation of individuals in the work force are essential to organizational success. A critical need for all these personnel functions is the identification of the work requirements of the job and the translation of these requirements into measurable criteria that will allow an assessment of the individual's current or potential proficiency. The focus of this research is criterion development, specifically, the development and assessment of a measurement methodology for validating selection systems, evaluating training programs, and identifying individual training deficiencies.

Over the years, a variety of measurement techniques have been used to assess job proficiency. They range from subjective to objective, and from general to specific. Historically, the emphasis has been on the selection of the "most available" criterion rather than on the development of the most appropriate criterion (Ronan & Prien, 1966).

The chief dilemma facing researchers in the area of criterion development has been aptly referred to as the "criterion problem" (Wallace & Weitz, 1955). At the heart of this problem lies the "ultimate" criterion which is the hypothetical true domain of job performance. The difficulty is in how to translate this hypothetical concept into a quantifiable and objective unit, free from the constraints of measurement and human judgment. Because this is impossible , the objective is to identify an actual or "immediate" criterion that is as close to the ultimate criterion as possible (Thorndike, 1949). Several "criteria for criteria" are important in determining the usefulness of any particular criterion measure. These include the concepts of relevance, deficiency, contamination, and redundancy.

Work Sample Testing

In the past, researchers (e.g., Ghiselli & Brown, 1948; Guion, 1979; Robertson & Kandola, 1982; Siegel, 1982) have touted work sample tests as the single-most direct, relevant measure of job proficiency. Work sample tests measure an individual's skill level by extracting samples of behavior under realistic job conditions. Guion (1979) discusses work samples in terms of two broad classes, direct and indirect/abstract. A direct work sample requires an individual to perform certain tasks specific to the job in question, whereas an abstract work sample identifies underlying skills required to perform a job and creates tests to assess the presence of those skills.

In the early 1980s requests from the Department of Defense and Air Force manpower, personnel, and training communities, as well as internal project requirements within the Air Force Human Resources Laboratory

(AFHRL) prompted the AFHRL to initiate the Job Performance Measurement (JPM) Project. The aim of this endeavor was to develop and test a measurement technology for assessing the job performance of enlisted personnel in their first 4 years of military service. Because the job requirements of these individuals are primarily technical in nature, the assessment of an individual's skill or job proficiency was identified as the criterion of interest (or ultimate criterion). Hands-on work sample testing was designated the most direct measure of task-level proficiency. Consequently, it became a benchmark measure against which other measures (e.g., rating forms, job knowledge tests, Interview Testing) could be compared. A brief review of previous work sample test development efforts follows, with a focus on jobs with a technical orientation.

Personnel Selection

Work sample tests have been used for many years as a method for selecting applicants into the workforce. As such, they have been designed primarily to assess present skill levels. As early as 1913, Munsterberg cited the development of a work sample test for selecting streetcar operators. In a review of the validity of work sample tests, Asher and Sciarrino (1974) provide examples of over 80 different tests developed between 1937 and 1972 for selecting job applicants. A more recent review of the personnel selection literature (Robertson & Kandola, 1982) also cites numerous instances of work sample testing, and reports a high percentage of predictive validities in the .40 to .60 range.

A second use of work sample testing in the selection domain is exemplified by the research of Robertson and Downs (Downs, 1970; Robertson & Downs, 1979, 1989) and Siegel (Siegel, 1982; Siegel & Bergman, 1975). This approach, referred to as trainability testing, or miniature job training and evaluation, focuses on identifying an individual's potential for training prior to being placed on the job. Job applicants are given short training sessions followed by testing sessions that assess what has been learned. The success of this approach has been reported most recently by Robertson and Kandola (1982) and Robertson and Downs (1989). These researchers note that trainability tests, like more traditional work sample tests, display high content validity and face validity. The fact that the test content is directly related to the job may ensure that applicants and assessors view them more favorably. Schmidt, Greenthal, Hunter, Berner, and Seaton (1977) also note the more positive attitudes of applicants to work sample tests when compared to written selection tests.

Job Proficiency

Whereas personnel selection has been the primary reason for work sample test development, work sample tests have also been used as criteria. In addition to assessing job proficiency for use as a criterion for validating a selection device (Siegel & Jensen, 1955), the approach has also been used to evaluate training programs (Goldstein, 1974), identify individual skill deficiencies (Goldstein, 1980), and establish worker job certification (Guion, 1979).

Because of the relative ease of development and adaptability to different jobs, rating forms are the most frequently used criteria for evaluating selection systems. Still, work sample tests have been used periodically for this purpose (Siegel & Jensen, 1955). In fact, researchers such as Wernimont and Campbell (1968) and Asher and Sciarrino (1974) call for more frequent use of work sample tests as criteria when validating selection systems.

Decisions concerning the worth of training programs require a mechanism for determining that the instructional program was responsible for the changes that occurred. Wilson (1962) discusses the conceptual basis for using work sample testing for training program evaluation, cites examples of its use, and concludes that training program evaluation is one of the best uses of the work sample measurement technique. Similarly, in a recent review of maintenance training research, Gibson and Orlansky (1986) identified two of seventeen studies that use work sample measurement to assess training effectiveness, and call for increased use of this measurement approach.

Work sample tests are also used to identify individual strengths and weaknesses in order to determine the skills and knowledges needed to perform the job successfully (Wexley & Latham, 1981). Just as with other criterion research, supervisory ratings are the most frequently used measure, but researchers continue to recommend use of work sample measures (Goldstein, 1980; Guion, 1979).

Finally, work sample testing has been used to certify that an individual meets or exceeds a designated level of competence at performing a job. The Army system of skill qualification testing (Maier, Young, & Hirshfeld , 1976) is an example of such a system, whereby a work sample approach is used to assess task competence on various components of a soldier's job. In another study, Fullerton, Peele, Reed, Morrison, and Liebowitz (1982) developed an innovative approach to certification as part of a licensing examination for nuclear power plant operators. An oral "walk-through" examination was designed such that an examiner questions and observes an individual while touring the plant and control room in an "operating demonstration." The examiner asks questions and makes notes of the answers and/or their adequacy. This oral examination, in conjunction with a work simulation and a written examination, formed the basis for the licensing examination.

Overview of Air Force Work Sample Testing

Hands-on testing presents a particular problem for the Air Force because of the complexity and expense involved in performing many tasks. For example, many critical tasks cannot be measured by hands-on testing because these tasks may take too long to complete, require replacement of expensive parts, and risk possible injury to personnel or damage to equipment. The AFHRL has developed a new methodology to address these problems. This new approach, Walk-Through Performance Testing (WTPT), has as its foundation the work sample philosophy, but expands the measurement of critical tasks to include those tasks not measured by hands-on testing through the use of an interview testing component (Hedge & Teachout, 1986).

Walk-Through Performance Testing

WTPT[1] is a task-level job performance measurement system that expands the range of job tasks on which an individual is measured by combining hands-on task performance and interview procedures to provide a measure of individual technical job competence. The interview testing component has been added as a means of assessing those critical tasks eliminated from the content domain during developmental efforts because of measurement constraints. The WTPT is a detailed step-by-step checklist which specifies the behavior, conditions, and standards for successful task accomplishment. It is written in technical terms and employs a dichotomous rating scale (i.e., yes/no) to record the occurrence of correct or incorrect performance on each step.

The hands-on component resembles a traditional hands-on work sample test designed to measure proficiency on a set of critical tasks. As the incumbent performs a task, the test administrator observes the performance and, using the WTPT as a checklist, records correct or incorrect completion of each step.

Interview Testing allows the administrator to measure proficiency on tasks precluded from hands-on measurement (i.e., tasks that are either too time-consuming, too costly, or too dangerous for hands-on measurement). Administration of the interview component of the WTPT differs from the hands-on component in that it requires the test administrator to direct the incumbent to describe in detail how he/she would perform a job-related task. The administrator assesses an incumbent's proficiency on a task by asking questions designed to uncover proficiency-based strengths and weaknesses related to the performance of that task. Again, using the WTPT as a checklist, the test administrator records correct or incorrect description of each step of the task. Figure 1 shows an example of an interview item.

Only one reference was found within the technical job domain literature to guide the conception and development of the Interview Testing approach. The work conducted by the Oak Ridge National Laboratory on their oral exam for nuclear power plant operators was the only work sample approach found to be conceptually similar in intent, namely, verbal description of steps required for successful task completion. As noted by Fullerton et al. (1982), preparation for the oral examination varies widely across examiners; some examiners develop multi-page oral forms of questions and anticipated answers, while others prepare lists of brief topics to discuss. Conduct of the examination also varies widely and is left to the discretion of the examiner. In their study, no official standards, definitions, or guidance was provided for what constitutes a satisfactory/marginal/unsatisfactory grading format. Summation of these scores into an overall rating is also left to the discretion of the examiner.

---

[1]WTPT may refer interchangeably to the Walk-Through Performance Test, the instrument, or to Walk-Through Performance Testing, the process.

Phase I Interview Task 233

Objective: To evaluate the incumbent's knowledge of procedures required to safety wire system components.

Estimated Time: 5M Start: Finish: Time Req:

Time Limit: 10M #Times Performed: Last Performed:

Tools and Equipment: .032 lockwire, lockwire trainer, lockwire pliers.

Background: A lockwire trainer was fabricated to provide standardization across MAJCOMs.

Configuration: Existing lockwire should be removed from the trainer.

Instructions to Administrator:

Administer at the interview table allowing the incumbent to look at the lockwire trainer.

SAY TO THE INCUMBENT

TELL ME THE STEP BY STEP PROCEDURES YOU WOULD FOLLOW TO SAFETY WIRE THE TWO WING NUTS IN ACCORDANCE WITH THE GENERAL LOCKWIRE PROCEDURES. REMEMBER TO DESCRIBE THIS TASK IN AS MUCH DETAIL AS POSSIBLE.

Performed or Answered Correctly	Yes	No
Did the incumbent say he/she would:		
1. Cut a length of lockwire approximately 18 inches long from the spool?	___	___
2. Select the hole in the uppermost wing of the left wing nut?	___	___
3. Feed one end of the safety wire through the hole in the left wing nut and pull approximately halfway through?	___	___
4. Measure the double strand of wire over the top of the left wing nut and under the right wing nut to the hole in the lower wing (tightening direction)?	___	___
5. Apply the pliers to the measured point on the double strand of lockwire and twist at a rate of 8 to 10 turns per inch?	___	___
6. Feed one end of the untwisted strand of wire through the selected hole in the right wing nut?	___	___
7. Check the twisted wire for proper tension?	___	___
8. Apply pliers 1 to 2 inches beyond the right wing nut and twist the double strand of wire?	___	___
9. Dike off the twisted wire 4 to 5 turns beyond the right wing nut?	___	___
10. Turn the pigtail into the wing nut so as to eliminate any hazard?	___	___
11. Test final assembly for proper tension and direction?	___	___

Figure 1. Sample Interview Item.

In contrast, AFHRL's Interview Testing adopts a much more rigorous approach to test development, clear definition of standards for correct performance, administrator training, test administration procedures, and scoring of performance. Incumbents are tested on a predetermined set of tasks, with a specific set of steps required for task completion. Administrators are trained to score performance on a dichotomous (correct/incorrect) format against predetermined standards required for correct task performance.

Objective of the Study

Hands-on work sample testing has a well-documented history of development and application, while the viability of the interview testing format is relatively unknown. The objective of this study was to compare Hands-on and Interview Testing in order to determine the viability of Interview Testing as a measurement methodology. Three questions were addressed to make this comparison:

1. Are there mean differences in test scores between hands-on and interview methods?

This indicates whether different inferences would be made about the proficiency of individuals.

2. What is the correlation between hands-on and interview methods?

This indicates whether the different methods agree with respect to the ordering of individuals.

3. Do hands-on and interview methods have similar or different patterns of relationships when each is correlated with other performance-relevant variables?

This provides additional evidence about whether the two methods tap similar constructs of behavior.

II. METHOD

The JPM project was initiated to develop a variety of job performance measures for use in the validation of selection/classification methodologies and evaluation of training programs. The AFHRL developed an Interview Testing approach to work sample measurement, in addition to the more traditional hands-on work sample approach. Over a five-year period, instruments were developed and data were collected on eight enlisted Air Force Specialties (AFSs).

Participants

Personnel from eight specialties[2] were included in this study: Aircrew Life Support Specialist (AFS 122X0); Air Traffic Control Operator (AFS 272X0); Precision Measurement Equipment Laboratory (PMEL) Specialist (AFS 324X0); Avionic Communications Specialist (AFS 328X0); Aerospace Ground Equipment (AGE) Mechanic (AFS 423X5); Jet Engine Mechanic (AFS 426X2); Information Systems Radio Operator (AFS 492X1); and Personnel Specialist (AFS 732X0). A total of 1491 job incumbents in their first enlistment (i.e., first 4-year commitment) were participants. The majority of the data collection took place at Air Force bases within the continental U.S., although Radio Operator personnel were tested world-wide. Table 1 presents descriptive statistics of sample characteristics.

Table 1. Sample Characteristics for Eight AFSs

Aptitude Grouping	Incumbents N	Months In Service	Gender % Male	Race % Caucasian
Mechanical				
AGE	261	28.4	92.0	87.7
Jet Engine	255	31.1	96.1	84.3
Administrative				
Radio Operator	156	22.8	61.8	59.9
Personnel	197	28.0	55.3	59.4
General				
Life Support	195	29.3	83.1	67.2
Air Traffic Control	191	26.7	88.5	85.3
Electronic				
PMEL	138	27.5	91.3	89.9
Avionic Comm	98	34.8	94.9	93.9

[2]Four aptitude indices (AI) from the Armed Services Vocational Aptitude Battery (ASVAB) are used in the classification of airmen. These groupings are Mechanical, Administrative, General, and Electronic. For facilitation of discussion and display purposes, the data from the eight specialties are grouped according to the appropriate AI for each specialty (see Table 1).

Measures

Walk-Through Performance Test

Each WTPT contained detailed step-by-step checklists which specified the conditions, standards, and behaviors for successful performance on a set of tasks representative of the job of the first term enlistee. WTPTs contained two work sample testing approaches, hands-on and interview; both approaches required the examinee to perform the tasks at the work setting under the observation of a test administrator who scored each step on a correct/incorrect basis. In the hands-on portion, the incumbents were instructed to perform each task according to technical order (TO) procedures. Examinees were allowed access to TOs or other written information necessary for task completion. Interview Testing required the incumbent to describe the steps necessary for task completion in a "show-and-tell" manner without the aid of technical information. Every task in a WTPT, 20 to 30 for each specialty, had a maximum time limit at which point task performance was stopped and all steps not performed were recorded as incorrect.[3]

WTPT Scoring Procedures. Each WTPT task was composed of a series of steps that had been previously weighted on a nine-point adjectivally anchored rating scale based on the importance of that step to the successful completion of the task. These weights were assigned by senior non-commissioned officers (NCOs) from each AFS during scoring workshops held prior to data collection. Weights were then summed across all steps for a task, creating a "base score" for that task. Weights for each step scored correctly performed were summed, divided by the base score, and multiplied by 10. This placed each task score on a 0-10 point scale, so that all tasks, regardless of the number of steps, received equal weight in the computation of the total test score. Task scores could then be summed to form composite hands-on, interview, and total WTPT scores.

Each WTPT contained a common set of tasks performed by a majority of the first term members of that AFS. These tasks were administered to all incumbents tested in a specialty. In addition, seven of the eight AFSs contained sets of tasks that were specific to groups of incumbents within an AFS, dependent on equipment and mission differences. [Note. Only AFS 423X5 WTPT consisted entirely of tasks common across the specialty.] Since inferences from test scores should be based on individual differences, not test differences, a procedure was used to adjust the scores on the specific sets of tasks to account for possible differences in test difficulty across sets of tasks. This adjustment involved converting the specific task scores to the same metric as the common tasks, so that scores would be equivalent. The Design III equating procedure described by Angoff (1982, pp. 1350-1351) was used to adjust the scores.

---

[3]Details of the AFHRL's domain sampling strategy used in this research can be found in Lipscomb and Dickinson (1988).

Other Measures

Rating, job knowledge, experience, and archival data were also gathered from job incumbents, supervisors, or coworkers. A series of four rating forms (Task, Dimensional, Air Force-wide, and Global) were developed to measure performance from the very specific (i.e., task-level) to the very general (i.e., Global). Ratings on each form were made (by supervisors, peers, and job incumbents) on a five-part adjectivally anchored scale.

Paper-and-pencil tests of procedural job knowledge were developed for four specialties (Life Support, PMEL, AGE, and Personnel) and administered to job incumbents, as was a job experience questionnaire.

Finally, archival data from personnel files were accessed. These incumbent data included time in service, technical training school final course grades, and ASVAB subtest scores.

Test Administrator Training

The work sample tests were administered to job incumbents by active-duty NCOs and former enlistees (for Air Traffic Control and Jet Engine specialties only) in the career fields tested. These test administrators (TAs) received 1 to 2 weeks of observation and scorer training (Hedge, Lipscomb, & Teachout, 1988). Training of the administrators included instruction in observation/evaluation, interviewing, and WTPT administration procedures. Methods of training consisted of lecture/discussion, exercises, role playing, and review of videotaped task performance. This type of training produced accurate and reliable scoring by Jet Engine test administrators. Hedge, Dickinson, and Bierstedt (1988) calculated scorer agreement and correlational accuracy indices between TA scores and videotape target scores. Reported average scorer agreement ($r = .81$) and accuracy ($r = .85$) were quite high.

In the remaining seven AFSs, videotapes of work sample test performance with known target scores were also used as a training/evaluation device. After viewing and scoring the videotapes, the trainees engaged in detailed discussions to identify key behaviors that an incumbent should perform and avoid for successful task completion. High levels of administrator reliability and accuracy were obtained for all specialties prior to test administration.

During data collection for the last seven specialties (excluding Jet Engine) a technique referred to as "shadow scoring" was used to evaluate test administrator agreement and facilitate retraining when necessary. In shadow scoring, two test administrators independently observe and score an individual performing a task. These scores are then compared and discussed to identify and resolve any discrepancies. Shadow scoring evaluation and retraining was effective in maintaining test administrator agreement in the scoring process. Across the seven specialities, test administrator agreement ranged from .92 to .98 (median $r = .97$) for Hands-on Tests and from .92 to .97 (median $r = .93$) for Interview Tests.

Procedure

In a group session, raters were introduced to the research project, participation conditions were explained, and they were familiarized with each measure used in the project. This orientation was followed by 1 hour of frame-of-reference and rater error training. Immediately following rater training, the job incumbents, peers, and supervisors completed the series of rating forms and associated questionnaires. Next, job incumbents in four AFSs (Life Support, PMEL, AGE, and Personnel) were group administered job knowledge tests.

The final testing stage, the WTPT, occurred over several days at each site. Each incumbent was tested individually by a TA. Administration required 4 to 8 hours per incumbent, with test length dependent on the specialty.

Performance was measured by the hands-on methodology only, the interview methodology only, or both methodologies. In all cases where items were constructed for a task by both approaches, the "overlap" interview item was administered prior to the hands-on item to minimize the transfer of performance from one format to another (i.e., hands-on performance utilizing TOs might facilitate subsequent interview proficiency but the reverse is unlikely).

Data Analysis Variables

Variables included in the analyses are displayed in Table 2. The analyses consisted of descriptive statistics for all study variables, mean differences between work sample tests, correlational analyses, tests of significance between work sample tests and other variables, task-level mean differences, and task-level correlational analyses. Pairwise deletion was used in all analyses.

III. RESULTS

Hands-on and interview work sample test scores were compared for eight specialties. Analyses were conducted at the test level and task level. Test-level analyses focused on aggregate test scores across hands-on and interview tasks. Analyses were conducted to determine mean differences between hands-on and interview tests, correlations between hands-on and interview tests, and correlations of hands-on and interview tests with performance-relevant variables. Task-level analyses focused on individual tasks that were measured by both the hands-on and interview methods (overlap tasks). Analyses were conducted to determine mean differences and correlations between hands-on and interview overlap tasks. The results of these analyses are described in the following sections.

Table 2. Variables Included in Analyses

WTPT Performance Scores

- Interview Test
  - All tasks
  - Overlap interview tasks
- Hands-on Test
  - All tasks
  - Overlap hands-on tasks

JPMS Performance Measures[a]

- Job Knowledge Test (JKT) total scores on four specialties
- Ratings on Task, Dimensional, Air Force-wide, and Global Rating Forms for three sources (Self, Supervisor, and Peer)

Experience Measures

- Technical training school final grade
- Time in Service
  - Total Active Federal Military Service (TAFMS)
- Task experience ratings of prior experience on WTPT tasks
- Mean number of times each WTPT task had been performed
- Mean length of time since tasks had last been performed

Aptitude Measures

- Armed Service Vocational Aptitude Battery (ASVAB)
  - Mean sum of standard scores of four composites:
    - Mechanical
    - Administrative
    - General
    - Electronic
- Armed Forces Qualifying Test (New AFQT)

[a]Composites were formed for Task, Dimensional, and Air Force-wide rating forms. Mean ratings for each source across all tasks in the Task Rating Form and across all dimensions in the Dimensional Rating Form were utilized for the analyses. Seven items from the Air Force-wide Rating Form were averaged to get a summary interpersonal rating while the single-item rating of technical performance was used.

Test-Level Analyses

Hands-on and interview work sample tests were compared at the overall test level (i.e., all hands-on and interview items). Mean differences and correlational relationships were assessed for each of the eight specialties.

Work Sample Test Score Mean Differences

Mean scores (total percent correct) for hands-on and interview tests (i.e., all tasks) were calculated for each of the eight specialties. Dependent t-tests were computed for each hands-on/interview matched pair to determine if the two methods provided the same information about the proficiency level of individuals. Table 3 shows that significant differences were found for six of the eight specialties. For five of these six specialties (Personnel was the exception) scores reflected a significantly higher proficiency level for hands-on scores than interview scores.

Table 3. Work Sample Test Scores: Descriptive Statistics and Tests of Significance Between Means

AFS	Hands-on (Mean)	SD	Interview (Mean)	SD	t-test
Mechanical					
AGE	57.32	9.47	46.85	10.01	*
Jet Engine	73.01	10.53	62.36	12.53	*
Administrative					
Radio Operator	75.65	14.72	75.43	17.82	NS
Personnel	74.23	13.19	79.65	11.58	*
General					
Life Support	70.72	13.57	59.33	15.93	*
Air Traffic Control	69.49	11.06	66.10	11.54	*
Electronic					
PMEL	77.26	8.57	76.63	9.36	NS
Avionic Comm	78.26	12.06	68.66	12.41	*

NS = Not significant.
*$p < .001$.

Work Sample Test Score Intercorrelations

Hands-on and interview test scores were intercorrelated across each of the eight specialties to determine if the two criterion measures ordered the performance of individuals similarly. These values are displayed in Table 4. Work sample correlations were computed on the summary composite scores for all items,[4] and overlap items. For all specialties, all work sample intercorrelations were found to be significant. The correlations between all hands-on and all interview task scores were moderate to high, ranging from .457 (PMEL) to .839 (Personnel), with a median correlation of .682.

Table 4. Correlations Among Hands-On and Interview Composite Scores for Eight Specialties

	WTPT Composite	
ASVAB Composite	All Items	Overlap Items
Mechanical		
AGE	.702	.716
Jet Engine	.567	.451
Administrative		
Radio Operator	.800	.843
Personnel	.839	.943
General		
Life Support	.591	.594
Air Traffic Control	.808	.822
Electronic		
PMEL	.457	.339
Avionic Communications	.662	.508

Note. All correlations are significant ($p < .001$).

[4]The term "item" refers to a specific WTPT test component; "task" refers to a specific job element for which a hands-on item, an interview item, or both were developed.

Relationships to Relevant Variables

Another step in examination of Interview Testing's relationship with the hands-on work sample was to assess how similarly these work sample measures related to other performance-relevant variables. Twenty-eight rating form, aptitude, and experience variables were included in analyses for Jet Engine, Radio Operator, Air Traffic Control, and Avionic Communications specialties. Twenty-nine variables (JKTs were included for the last four specialties) were included in analyses for AGE, Personnel, Life Support, and PMEL. Descriptive statistics for these variables can be found in Appendix A.

Each of these performance-relevant variables was correlated with scores for (a) all hands-on items and (b) all interview items. A two-tailed test of significance[5] was computed for each pair of correlations. Table 5 depicts the presence of all significant differences across the eight specialties. A significant difference in magnitude of the correlation between the independent measure and the two criterion scores is an undesirable finding if you are trying to establish criterion comparability. These analyses identified relatively few (7.89%, 18 of 228) significant differences (critical $Z = 1.96$, $p < .05$). This lends additional support to the similarity between the interview and hands-on methodologies. (The correlational values between the work sample measures and performance-relevant factors for each specialty are contained in Appendix B.)

Task-Level Analyses

To examine the two work sample measures more closely, task-level analyses were performed on all overlap task means. In addition, correlational analyses were performed on all overlap tasks for all eight specialties.

Work Sample Task Score Differences

Mean task scores for each hands-on and interview overlap item were computed for each of the eight specialties. (Descriptive statistics for all hands-on and interview items are contained in Appendix C). Differences between these mean values were then assessed using dependent t-tests for each task. Tables 6, 7, 8, and 9 present a summary of these findings across the eight specialties.

---

[5]See Roscoe (1975, p. 267-268) for tests of significance between two Pearson correlation coefficients from related samples.

Table 5. Significant Differences Between Work Sample Measures and Other Relevant Variables

	Specialties							
	Mechanical		Administrative		General		Electronic	
Variable	AGE	JEM	ISRO	PERS	ALS	ATC	PMEL	ACS
RATINGS								
Task					*			
Self						*	*	
Super								
Peer								
Dimensional								
Self							*	
Super								
Peer								
Global-Technical								
Self							*	
Super								
Peer	*							
Global-Interpersonal								
Self								
Super								
Peer								
AFW Tech								
Self						*	*	
Super								
Peer		*						
AFW Inter								
Self	*							
Super								
Peer								
Knowledge Test					*			
EXPERIENCE								
Times Performed					*			
Last Performed (Wks)					*	*		
Experience Rating					*			
Months in Service								
Training Grade								
APTITUDE								
Mechanical						*		
Administrative								
General								
Electronic	*	*						
New AFQT								

Note. An asterisk designates a significant difference ($p < .05$) between correlations of performance-relevant variables to hands-on and interview work samples. Job Knowledge Tests were developed for four specialties only (AGE, Personnel, Life Support, and PMEL).

Table 6. Task-Level Summary Table for AGE and Jet Engine Overlap Tasks

Task[a]	N	Interview Mean	Hands-on Mean	Correlation
AGE				
251	261	1.34	2.81*	.64
209	261	5.92	6.50*	.65
300	261	6.05	7.58*	.53
264	261	4.26	5.19*	.51
421	261	2.98	3.14	.73
Jet Engine				
134	255	6.15	7.21*	.34
374	255	6.67	7.02*	.35
373	255	6.26	8.48*	.19
353	255	6.52	7.62*	.23
360	255	5.93	7.18*	.39

Note. All correlations are significant ($p < .05$).

[a]Numbers represent tasks for which both hands-on and interview items were developed.

*Indicates a mean that is significantly ($p < .05$) greater than the other in the pair.

Table 7. Task-Level Summary Table for Radio Operator and Personnel Overlap Tasks

Task[a]	N	Interview Mean	Hands-on Mean	Correlation
Radio Operator				
218	157	7.96	7.72	.84
209	157	5.25	8.17*	.37
201	157	9.05	9.62*	.60
142	157	8.81	8.85	.62
173	68	3.83	3.78	.88
184	68	7.94	8.49	.72
166	48	7.93	8.40*	.72
197	48	7.96	8.11	.46
183	41	8.23	8.21	.78
197	41	8.08	8.33	.80
Personnel				
035	197	7.12	7.38*	.92
719	197	9.18	9.17	.87
140	197	7.38	7.24	.92
145	31	8.81*	7.92	.81
293	31	5.42	6.18	.69
334	31	4.27	4.80	.90
476	38	7.87	8.09	.90
001	38	8.04	8.15	.83
466	38	8.08	8.76*	.67
415	46	8.22	8.34	.88
396	46	8.56	8.80*	.95
380	46	6.01	5.94	.73
728	35	8.36	8.59	.73
835	35	9.08	8.92	.63
874	35	5.76	5.48	.97
876	47	5.00	4.87	.92

Note. All correlations are significant ($p < .05$).

[a]Numbers represent tasks for which both a hands-on and interview item were developed.

*Indicates a mean that is significantly ($p < .05$) greater than the other in the pair.

Table 8. Task-Level Summary Table for Life Support and Air Traffic Control Overlap Tasks

Task[a]	N	Interview Mean	Hands-on Mean	Correlation
Life Support				
330	195	6.10	6.92*	.52
295	195	3.40	6.07*	.46
320	195	7.86	9.48*	.24
315	195	6.74	7.60*	.41
389	195	6.42	8.11*	.43
383	195	5.07	6.91*	.52
Air Traffic Control				
172	191	6.29*	5.87	.49
274	191	7.63	7.86	.65
293	191	8.88*	8.12	.36
406	52	8.06	8.83*	.68
405	52	7.27	7.24	.73
381	52	4.72	4.73	.85
232	139	5.13	5.16	.85
271	139	4.67	5.06*	.83
369	139	7.28	7.84*	.90

Note. All correlations are significant ($p < .05$).

[a]Numbers represent tasks for which both a hands-on and interview item were developed.

*Indicates a mean that is significantly ($p < .05$) greater than the other in the pair.

Table 9. Task-Level Summary Table for PMEL and Avionic Communications Overlap Tasks

Task[a]	N	Interview Mean	Hands-on Mean	Correlation
PMEL				
211	138	8.88	9.38*	.30
214	138	7.52	8.17*	.39
239	138	7.77	9.17*	.23
403	138	8.64	9.93*	.48
336	138	5.77	7.27*	.30
Avionic Comm				
160	98	8.56	9.12*	.48
232	98	6.48	7.98*	.40
233	98	8.15	9.31*	.28
234	98	5.58	6.42*	.82
257	98	7.66	8.92*	.57
458	35	4.98	6.87*	.77
459	32	5.84	9.65*	.43
258	31	4.57	4.42	.98

Note. All correlations are significant ($p < .05$).

[a]Numbers represent tasks for which both a hands-on and interview item were developed.

*Indicates a mean that is significantly ($p < .05$) greater than the other in the pair.

Significant hands-on and interview mean differences were found on 61% of all overlap task pairs. Significant differences were most notable in the Electronic (92%) and Mechanical (90%) career fields. In addition, across all specialties, hands-on scores were higher than interview scores in a relatively large percentage of items (78%). Once again, this trend was most pronounced for the Electronic (92%) and Mechanical (100%) career fields. These findings can also be found in Tables 6 through 9. While Electronic, Mechanical, and General career fields all showed a preponderance of larger hands-on scores, this was not the case in the Administrative career fields, where only 27% of all comparisons showed larger hands-on means.

Work Sample Overlap Task Intercorrelations

Hands-on and interview overlap task scores were intercorrelated within each AFS to provide a closer look at the viability of Interview Testing. Correlations for all overlap items across the eight specialties are reported in Tables 6, 7, 8, and 9. Correlations ranged from .19 to .97 across the eight specialties, with the greatest number of larger correlations found in Administrative specialties (i.e., Radio Operator and Personnel). The correlations for these specialties ranged from .37 to .97, with a median correlation of .81. Median correlations for the specialties in the remaining three AI areas were .52 (General), .45 (Mechanical), and .42 (Electronic).

IV. DISCUSSION

The purpose of this study was to examine Interview Testing as a viable work sample methodology. This was done by comparing the interview approach with the more traditional hands-on work sample approach. Mean differences and correlational differences between these two methodologies were investigated at both the overall test level and the individual task level. Correlational analysis examines comparable ordering of individuals by the two methodologies. Analysis of mean differences indicates whether individuals score at similar levels on the two work sample tests. Both sets of analyses provide insights into the viability of the interview as a work sample methodology, especially when applied to different personnel functions (e.g., selection system validation, or training needs assessment).

Correlational Analyses

Test-level analyses generated moderate to high correlations between hands-on and interview work sample tests across the eight Air Force specialties (.439 to .839), with a median correlation of .682. This suggests a relatively similar ordering of individuals by the two methodologies. In addition, tests of differences between the two work sample measures and other relevant variables (e.g., ratings, aptitude, experience) found few correlational differences. In fact, out of 228 predictor-criterion combinations, only 18 significant differences (i.e., the size of the hands-on/predictor correlation was significantly different

from the size of the interview/predictor correlation) were found. Thus, the hands-on and interview work samples correlate similarly with other variables in over 92% of the cases across the eight specialties.

As noted by Green (1984), to determine the comparability of two techniques, it should be established that they are both measuring essentially the same construct. This was done by examining both criterion intercorrelations and the similarity of these criterion relationships to other relevant variables. These test-level results suggest that Interview Testing shows considerable promise as a work sample measurement methodology.

Task-level correlational analyses performed on each task having both an interview and hands-on item provided less consistent results across the eight specialties. In all, 64 hands-on/interview overlap items were analyzed. Correlations ranged from .19 to .97 across the eight specialties. Magnitude of the hands-on/interview correlation appears to be a function (at least in part) of the aptitude index area under investigation. Task-level correlations in the two mechanical specialties range from .19 to .73, while correlations range from .46 to .97 in Administrative specialties, .24 to .90 in General specialties, and .23 to .82 in Electronic specialties. Overall, 65% of the task-level correlations are .50 or larger. Only 25% of the Electronic overlap items correlate .50 or larger, while over 73% of the Administrative correlations are .70 or larger. These findings suggest that the researcher can place less confidence in a hands-on/interview comparability assumption when operating at the task level.

Mean Differences

Test-level analyses identified significant mean differences between hands-on and interview test scores for six of the eight specialties. In addition, in all but one of these specialties, the hands-on percent correct score was larger than the interview test score. These results suggest that individuals score at different performance levels on the two work sample tests.

These findings were replicated with task-level analyses, where 39 of 64 (60.9%) mean comparisons (across the eight specialties) were significantly different. In addition, across the 64 comparisons, hands-on scores were larger than interview scores 78% of the time. These differences did vary, however, by aptitude index area. Significant mean differences were found in 92% of task comparisons in Electronic specialties, in 90% of Mechanical specialty tasks, in 93% of General specialty tasks, and in only 27% of Administrative specialty tasks. Interestingly, the trends across these specialties (as depicted in Tables 6, 7, 8, and 9) are for specialties with more concrete/motor tasks to have a greater number of hands-on tasks with larger means, while specialties with more abstract/verbal tasks have interview tasks with larger means.

While these mean differences may be due to the type of tasks (motor versus verbal), several alternate explanations are also feasible. For example, these differences could reflect differential abilities; thus less

verbally proficient individuals may be placed in specialties with lower verbal requirements. Another explanation may be simply that these interview tests are more difficult than the hands-on tests, in general, and this difficulty varies somewhat across specialties. In summary, though, a preponderance of mean differences between the two work sample approaches were found at both the test and task levels of analysis.

Implications for Personnel Decisions

Correlational and mean difference analyses at the test and task levels have suggested different conclusions about comparability of the two work sample methodologies. More importantly, these analyses provide different insights into the use of these two approaches for different personnel functions. A brief look at how comparability requirements differ by personnel function should help to clarify this notion.

Selection system and training program validation. The primary purpose of selection system validation is top-down selection of the same recruits, and the primary purpose of training program evaluation is to establish that how individuals perform in training reflects how they will perform on the job. Thus, to establish a comparability link between hands-on and interview work sample measures for validation purposes, your analyses must demonstrate a similar ordering of individual performance scores (high criterion intercorrelations). Similarity of performance levels (i.e., mean test scores) on the two criterion measures in no way affects the ordering of individuals. In terms of the current research results, a test-level (the appropriate level of analysis for these purposes) median correlation of .68 between Hands-on Testing and Interview Testing is relatively strong, suggesting Interview Testing shows considerable promise as a work sample methodology for validation purposes.

Certification and training needs assessment. Job certification, task certification, and training needs assessment all require utilization of both correlational analysis and an analysis of mean scores. For job certification purposes, a two stage process is required to examine criterion comparability. A first, necessary condition is that individuals are ordered similarly. Second, it must be determined whether individuals' performance scores on the two tests are similar. If both of these conditions are met, the two approaches can be considered comparable. If individuals are ordered differently, then the criteria cannot be considered comparable. If performance levels are different, separate competency standards must be established for the two work sample measures.

In terms of the present research findings, a test-level median correlation of .68 suggests a moderate confirmation of the first comparability requirement. However, there were significant mean differences between hands-on and interview test scores for six of the eight specialties. Thus, use of Interview Testing for the six specialties (with different performance scores) for certification or training needs assessment would require the establishment of different competency standards than those that exist for the hands-on tests.

These same principles that operate for job certification purposes must be applied to task certification and training needs assessment. However,

rather than operating at the test-level as is the case with job certification, training needs assessment and task certification address task-level competency issues. In terms of the present research findings, comparability conclusions are less clear-cut. Task-level correlations vary considerably both within and between specialties, ranging from .19 to .97. The work sample methodologies order individuals similarly in the Administrative specialties (median correlation = .81), but this impressive magnitude drops continually as you move from General (median correlation = .52) to Mechanical (median correlation = .45) to Electronic specialties (median correlation = .42). In addition, the percent of mean differences vary by aptitude area in a similar manner. Thus, except for the Administrative specialties, hands-on/interview comparability is questionable.

Summary

This research study has examined a new work sample measurement methodology, Interview Testing, in relation to the more traditional hands-on work sample approach. In addition, the results of these comparisons have been viewed in the context of different personnel functions. Interview Testing shows great promise as a work sample methodology when used for validation research. Its usefulness for training needs assessment and certification purposes is more tenuous due to the variability of correlational and mean values across specialties. Still, because of cost and time savings, and because it allows assessment of proficiency on tasks not measurable in a "hands-on" fashion, Interview Testing shows potential as a work sample measurement methodology.

REFERENCES

Angoff, W. H. (1982). Norms and scales. In H. E. Mitzel (Ed.), Encyclopedia of educational research (Vol. 3, 5th ed., pp. 1342-1354). New York: Free Press.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Downs, S. (1970). Predicting training potential. Personnel Management, 2, 26-28.

Fullerton, A. M., Peele, E., Reed, J. H., Morrison, G. W., & Liebowitz, S. J. (1982, December). Final report on the evaluation of the nuclear power plant operator licensing examination. Oak Ridge, TN: Oak Ridge National Laboratory.

Ghiselli, E., & Brown, C. W. (1948). Personnel and industrial psychology. New York: McGraw-Hill.

Gibson, R. S., & Orlansky, J. (1986, September). Performance measures for evaluating the effectiveness of maintenance training (IDA Paper P-1922). Alexandria, VA: Institute for Defense Analysis.

Goldstein, I. L. (1974). Training: Program development and evaluation. Monterey, CA: Brooks/Cole.

Goldstein, I. L. (1980). Training in work organizations. Annual Review of Psychology, 31, 229-272.

Green, B. (1984). Measure surrogates and the problem of substitutability. Unpublished manuscript.

Guion, R. M. (1979, April). Principles of work sample testing: I. A non-empirical taxonomy of test uses (ARI-TR-79-A8). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hedge, J. W., Dickinson, T. L., & Bierstedt, S. A. (1988, July). The use of videotape technology to train administrators of Walk-Through Performance Testing (AFHRL-TP-87-71, AD-A195 944). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hedge, J. W., Lipscomb, M. S., & Teachout, M. S. (1988, June). Work sample testing in the Air Force Job Performance Measurement Project. In M. S. Lipscomb & J. W. Hedge (Eds), Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel (AFHRL-TP-87-58, AD-A195 630). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hedge, J. W., & Teachout, M. S. (1986, November). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37, AD-A174 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Lipscomb, M. S., & Dickinson, T. L. (1988, June). The Air Force domain specification and sampling plan. In M. S. Lipscomb & J. W. Hedge (Eds.), Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel (AFHRL-TP-87-58. AD-A195 630). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Maier, M. H., Young, D. L., & Hirshfeld, S. F. (1976, April). Implementing the skill qualification testing system (ARI-TR-76-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Robertson, I. T., & Downs, S. (1979). Learning and the prediction of performance: Development of trainability testing in the United Kingdom. Journal of Applied Psychology, 64, 42-50.

Robertson, I. T., & Downs, S. (1989). Work sample tests of trainability: A meta-analysis. Journal of Applied Psychology, 74, 402-410.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. Journal of Occupational Psychology, 55, 171-183.

Ronan, W. W., & Prien, E. P. (1966, June). Toward a criterion theory: A review and analysis of research and opinion. Creativity Research Institute, Richardson Foundation.

Roscoe, J. T. (1975). Fundamental research statistics for the behavioral sciences. New York: Holt, Rinehart and Winston.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper and pencil trades and technical tests: Adverse impact and examinee attitudes. Personnel Psychology, 30, 187-197.

Siegel, A. I. (1982, November). Work sample and miniature job training and evaluation testing. Paper presented at a conference on Performance Assessment: The State of the Art. Baltimore: Johns Hopkins University.

Siegel, A. I., & Bergman, B. B. (1975). A job learning approach to performance prediction. Personnel Psychology, 28, 325-339.

Siegel, A. I., & Jensen, J. (1955). The development of a job sample troubleshooting performance examination. Journal of Applied Psychology, 39, 343-347.

Thorndike, R. L. (1949). Personnel selection: Test and measurement techniques. New York: Wiley.

Wallace, S. R., & Weitz, J. (1955). Industrial psychology. Annual Review of Psychology, 6, 217-250.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

Wexley, K. N., & Latham, G. P. (1981). Developing and training human resources in organizations. Glenview, IL: Scott, Foresman.

Wilson, C. L. (1962). On-the-job and operational criteria. In R. Glaser (Ed.), Training research and education, Pittsburgh: University of Pittsburgh Press.

APPENDIX A: DESCRIPTIVE STATISTICS FOR JPMS VARIABLES

APPENDIX A-1: DESCRIPTIVE STATISTICS FOR JPMS VARIABLES (AFS 423X5 AND AFS 426X2)

Variable	N	Mean	SD	Limits	Median
AFS 423X5					
Experience					
Number Times Performed[a]	261	26.99	29.01	0 - 999	16.44
Last Time Performed (Wks)[a]	261	19.78	14.39	0 - 208	15.89
Task Experience Rating[a]	259	2.97	.71	1 - 5	2.83
TAFMS (Months in Service)[b]	261	28.38	10.07	4 - 54	27.00
Technical Training Grade	255	89.42	4.84	70 - 99	90.00
JKT (% Correct)	261	60.10	7.58	0 - 100	61.01
Aptitude (ASVAB Composites)[c]					
Mechanical	219	230.25	17.44	0 - 320	231.00
Administrative	219	155.93	12.56	0 - 240	156.00
General	219	106.85	8.56	0 - 160	105.00
Electronic	219	216.22	17.52	0 - 320	213.00
New AFQT	219	212.67	17.01	0 - 320	210.00
AFS 426X2					
Experience					
Number Times Performed	255	90.35	51.17	0 - 999	83.13
Last Time Performed (Wks)	255	13.14	20.87	0 - 208	9.12
Task Experience Rating	255	4.47	1.13	1 - 5	4.40
TAFMS (Months in Service)[d]	239	31.11	11.97	10 - 70	31.00
Technical Training Grade	247	86.35	6.56	70 - 99	86.00
Aptitude (ASVAB Composites)					
Mechanical	201	231.58	20.97	0 - 320	235.00
Administrative	201	153.06	13.55	0 - 240	154.00
General	201	106.23	9.93	0 - 160	106.00
Electronic	201	215.35	20.53	0 - 320	213.00
New AFQT	201	209.76	20.04	0 - 320	208.00

[a]Mean of ratings across all WTPT tasks.
[b]Note that two incumbents fell outside the focal 6-48 TAFMS range.
[c]Mean of sum of standard scores.
[d]Note that 12 incumbents fell outside the focal 6-48 TAFMS range.

APPENDIX A-2: DESCRIPTIVE STATISTICS FOR JPMS VARIABLES (AFS 492X1 AND AFS 732X0)

Variable	N	Mean	SD	Limits	Median
AFS 492X1					
Experience					
Number Times Performed[a]	157	223.98	211.92	0 - 999	166.30
Last Time Performed (Wks)[a]	157	7.30	5.55	0 - 208	5.31
Task Experience Rating[a]	157	3.70	.83	1 - 5	3.75
TAFMS (Months in Service)[b]	157	22.77	12.84	6 - 50	18.00
Technical Training Grade	157	82.95	7.15	70 - 99	83.00
Aptitude (ASVAB Composites)[c]					
Mechanical	128	199.29	27.55	0 - 320	196.00
Administrative	128	166.95	9.19	0 - 240	166.00
General	128	105.27	9.75	0 - 160	105.00
Electronic	128	205.51	24.16	0 - 320	201.50
New AFQT	128	210.73	19.51	0 - 320	209.00
AFS 732X0					
Experience					
Number Times Performed	197	267.46	182.31	0 - 999	226.50
Last Time Performed (Wks)	197	6.73	6.85	0 - 208	4.27
Task Experience Rating	197	3.82	.65	1 - 5	3.89
TAFMS (Months in Service)[d]	197	28.04	11.53	6 - 68	27.00
Technical Training Grade	195	88.58	5.38	70 - 99	89.00
JKT (% Correct)	196	74.36	9.37	0 - 100	76.00
Aptitude (ASVAB Composites)					
Mechanical	179	195.35	25.62	0 - 320	193.00
Administrative	179	168.37	9.45	0 - 240	168.00
General	179	105.82	9.06	0 - 160	105.00
Electronic	179	204.82	21.07	0 - 320	202.00
New AFQT	179	212.35	18.04	0 - 320	209.00

[a]Mean of ratings across all WTPT tasks.
[b]Note that two incumbents fell outside the focal 6-48 TAFMS range.
[c]Mean of sum of standard scores.
[d]Note that four incumbents fell outside the focal 6-48 TAFMS range.

APPENDIX A-3: DESCRIPTIVE STATISTICS FOR JPMS VARIABLES (AFS 122X0 AND AFS 272X0)

Variable	N	Mean	SD	Limits	Median
AFS 122X0					
Experience					
Number Times Performed[a]	195	105.30	118.83	0 - 999	65.69
Last Time Performed (Wks)[a]	195	29.87	21.43	0 - 208	26.33
Task Experience Rating[a]	188	3.28	.79	1 - 5	3.38
TAFMS (Months in Service)[b]	195	29.26	10.96	6 - 50	31.00
Technical Training Grade	194	89.85	4.55	70 - 99	64.58
JKT (% Correct)	194	63.45	10.18		
Aptitude (ASVAB Composites)[c]					
Mechanical	172	210.51	26.86	0 - 320	208.50
Administrative	172	159.98	12.65	0 - 240	160.00
General	172	106.81	8.95	0 - 160	106.50
Electronic	172	211.54	20.04	0 - 320	208.00
New AFQT	172	212.95	17.97	0 - 320	211.00
AFS 272X0					
Experience					
Number Times Performed	191	253.21	153.89	0 - 999	360.70
Last Time Performed (Wks)	191	7.97	5.68	0 - 208	6.36
Task Experience Rating	191	3.50	.69	1 - 5	3.53
TAFMS (Months in Service)	191	26.71	8.86	6 - 48	26.00
Technical Training Grade	116	86.34	5.58	70 - 99	87.00
Aptitude (ASVAB Composites)					
Mechanical	172	226.23	24.70	0 - 320	229.50
Administrative	172	165.99	12.12	0 - 240	167.00
General	172	114.44	7.35	0 - 160	115.00
Electronic	172	226.58	18.67	0 - 320	228.00
New AFQT	172	227.26	15.35	0 - 320	226.00

[a]Mean of ratings across all WTPT tasks.
[b]Note that two incumbents fell outside the focal 6-48 TAFMS range.
[c]Mean of sum of standard scores.

APPENDIX A-4: DESCRIPTIVE STATISTICS FOR JPMS VARIABLES (AFS 324X0 AND AFS 328X0)

Variable	N	Mean	SD	Limits	Median
AFS 324X0					
Experience					
Number Times Performed[a]	138	68.00	58.72	0 - 999	45.53
Last Time Performed (Wks)[a]	138	15.62	10.87	0 - 208	13.79
Task Experience Rating[a]	137	3.10	.60	1 - 5	3.05
TAFMS (Months in Service)[b]	138	27.47	10.44	14 - 52	25.00
Technical Training Grade	138	88.21	5.22	70 - 99	88.00
JKT (% Correct)	138	60.92	9.59	0 - 100	61.11
Aptitude (ASVAB Composites)[c]					
Mechanical	126	239.94	18.89	0 - 320	243.00
Administrative	126	168.55	14.06	0 - 240	171.00
General	126	116.95	6.87	0 - 160	118.00
Electronic	126	240.82	13.29	0 - 320	240.50
New AFQT	126	234.21	13.78	0 - 320	236.50
AFS 328X0					
Experience					
Number Times Performed	98	115.66	87.98	0 - 999	95.12
Last Time Performed (Wks)	98	11.32	8.55	0 - 208	8.47
Task Experience Rating	98	3.39	.69	1 - 5	3.41
TAFMS (Months in Service)[d]	94	34.81	15.31	6 - 69	36.50
Technical Training Grade	97	91.18	4.80	70 - 99	92.00
Aptitude (ASVAB Composites)					
Mechanical	87	242.61	16.67	0 - 320	244.00
Administrative	87	166.38	12.13	0 - 240	169.00
General	87	117.13	7.09	0 - 160	118.00
Electronic	87	242.00	13.48	0 - 320	242.00
New AFQT	87	234.75	13.02	0 - 320	237.00

[a]Mean of ratings across all WTPT tasks.

[b]Note that two incumbents fell outside the focal 6-48 TAFMS range. The length of the AFS 324X0 technical school (i.e., 10 months) prevented testing of the lower end of the range.

[c]Mean of sum of standard scores.

[d]Small sample size necessitated expanding the range of testing to include 18 incumbents beyond the 48 month TAFMS limit.

APPENDIX B: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES

APPENDIX B-1: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 423X5)

	Hands-on Total	Interview Total
Number Times Performed[a]	.204**	.181*
Last Time Performed (Weeks)[a]	.143	.162*
Task Experience Ratings[a]	.249**	.245**
Task Ratings[a]		
- Self	.280**	.316**
- Supervisor	.292**	.281**
- Peer	.337**	.345**
Dimensional Ratings		
- Self	.300**	.348**
- Supervisor	.333**	.286**
- Peer	.326**	.345**
Global Ratings - Technical		
- Self	.342**	.357**
- Supervisor	.254**	.222**
- Peer	.328**	.375**
Global Ratings - Interpersonal		
- Self	.161*	.150*
- Supervisor	.126	.113
- Peer	.192**	.217**
AF-Wide Ratings - Technical		
- Self	.271**	.332**
- Supervisor	.272**	.257**
- Peer	.307**	.340**
AF-Wide Ratings - Interpersonal		
- Self	.144	.159*
- Supervisor	.190	.170
- Peer	.175*	.207**
ASVAB Composites		
Mechanical	.342**	.273**
Administrative	.022	.011
General	.160*	.074
Electronic	.310**	.188*
New AFQT	.177*	.078
TAFMS (Months in Service)	.253**	.300**
Technical Training Grade	.313**	.234**
Job Knowledge Test (% Correct)	.416**	.416**

Note. Self, N = 261; Supervisor, N = 259; Peers, N = 659.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX B-2: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 426X2)

	Hands-on Total	Interview Total
Number Times Performed[a]	.242**	.329**
Last Time Performed (Weeks)[a]	.103	.042
Task Experience Ratings[a]	.242**	.228**
Task Ratings[a]		
- Self	.213**	.249**
- Supervisor	.178*	.158*
- Peer	.109	.073
Dimensional Ratings		
- Self	.177*	.214**
- Supervisor	.259**	.173*
- Peer	.164*	.093
Global Ratings - Technical		
- Self	.198**	.151*
- Supervisor	.341**	.237**
- Peer	.314**	.204*
Global Ratings - Interpersonal		
- Self	.081	-.005
- Supervisor	.205**	.111
- Peer	.070	-.025
AF-Wide Ratings - Technical		
- Self	.216**	.242**
- Supervisor	.315**	.237**
- Peer	.300**	.191*
AF-Wide Ratings - Interpersonal		
- Self	.019	-.056
- Supervisor	.220**	.099
- Peer	.119	.055
ASVAB Composites		
Mechanical	.169*	.223**
Administrative	.087	.092
General	.110	.219**
Electronic	.122	.251**
New AFQT	.066	.192*
TAFMS (Months in Service)	.172*	.202**
Technical Training Grade	.249**	.225**

Note. Self, N = 255; Supervisor, N = 250; Peer, N = 226.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX B-3: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 492X1)

	Hands-on Total	Interview Total
Number Times Performed[a]	.256**	.274**
Last Time Performed (Weeks)[a]	.101	.146
Task Experience Ratings[a]	.176	.184
Task Ratings[a]		
- Self	.237*	.239*
- Supervisor	.300**	.278**
- Peer	.206*	.183
Dimensional Ratings		
- Self	.243*	.227*
- Supervisor	.319**	.268**
- Peer	.226*	.236*
Global Ratings - Technical		
- Self	.365**	.297**
- Supervisor	.269**	.352**
- Peer	.239*	.303**
Global Ratings - Interpersonal		
- Self	.018	-.053
- Supervisor	.145	.187
- Peer	.090	.107
AF-Wide Ratings - Technical		
- Self	.294**	.212*
- Supervisor	.259**	.342**
- Peer	.295**	.399**
AF-Wide Ratings - Interpersonal		
- Self	.043	.016
- Supervisor	.064	.090
- Peer	.082	.134
ASVAB Composites		
Mechanical	.190	.255*
Administrative	-.095	-.036
General	.318**	.334**
Electronic	.289**	.348**
New AFQT	.320**	.359**
TAFMS (Months in Service)	.338**	.320**
Technical Training Grade	.270*	.350**

Note. Self, N = 156; Supervisor, N = 151; Peer, N = 373.
[a]Mean of ratings across all WTPT tasks.
*p < .01.
**p < .001.

APPENDIX B-4: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 732X0)

	Hands-on Total	Interview Total
Number Times Performed[a]	.349**	.364**
Last Time Performed (Weeks)[a]	.085	.062
Task Experience Ratings[a]	.294**	.255**
Task Ratings[a]		
- Self	.276**	.252**
- Supervisor	.258**	.294**
- Peer	.367**	.311**
Dimensional Ratings		
- Self	.266**	.257**
- Supervisor	.289**	.299**
- Peer	.255**	.229**
Global Ratings - Technical		
- Self	.334**	.272**
- Supervisor	.242**	.259**
- Peer	.282**	.236**
Global Ratings - Interpersonal		
- Self	-.003	.042
- Supervisor	.119	.088
- Peer	.022	-.014
AF-Wide Ratings - Technical		
- Self	.179*	.174*
- Supervisor	.249**	.222**
- Peer	.273**	.274**
AF-Wide Ratings - Interpersonal		
- Self	-.014	.029
- Supervisor	.157	.143
- Peer	.075	.059
ASVAB Composites		
Mechanical	.133	.100
Administrative	.185*	.175*
General	.271**	.239**
Electronic	.189*	.150
New AFQT	.270**	.238**
TAFMS (Months in Service)	.350**	.371**
Technical Training Grade	.266**	.194*
Job Knowledge Test (% Correct)	.297**	.306**

Note. Self, N = 197; Supervisor, N = 195; Peer, N = 416.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX B-5: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 122X0)

	Hands-on Total	Interview Total
Number Times Performed[a]	.238**	.366**
Last Time Performed (Weeks)[a]	.041	-.208*
Task Experience Ratings[a]	.297**	.429**
Task Ratings[a]		
- Self	.213*	.336**
- Supervisor	.180*	.135
- Peer	.223**	.247**
Dimensional Ratings		
- Self	.222**	.287**
- Supervisor	.158	.148
- Peer	.181*	.231**
Global Ratings - Technical		
- Self	.166	.156
- Supervisor	.202*	.120
- Peer	.135	.128
Global Ratings - Interpersonal		
- Self	.032	.108
- Supervisor	.060	-.004
- Peer	.020	-.036
AF-Wide Ratings - Technical		
- Self	.107	.127
- Supervisor	.101	.044
- Peer	.173*	.110
AF-Wide Ratings - Interpersonal		
- Self	-.026	.010
- Supervisor	.071	.031
- Peer	.067	.030
ASVAB Composites		
Mechanical	.183*	.293**
Administrative	-.077	-.159
General	.116	.083
Electronic	.158	.169
New AFQT	.108	.083
TAFMS (Months in Service)	.315**	.280**
Technical Training Grade	.115	.106
Job Knowledge Test (% Correct)	.501**	.631**

Note. Self, N = 195; Supervisor, N = 189; Peer, N = 486.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX B-6: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 272X0)

	Hands-on Total	Interview Total
Number Times Performed[a]	.260**	.241**
Last Time Performed (Weeks)[a]	-.043	-.136
Task Experience Ratings[a]	.153	.148
Task Ratings[a]		
- Self	.177*	.192*
- Supervisor	.065	.080
- Peer	.207*	.207*
Dimensional Ratings		
- Self	.198*	.190*
- Supervisor	.074	.072
- Peer	.198*	.158
Global Ratings - Technical		
- Self	.200*	.202*
- Supervisor	.172*	.119
- Peer	.264**	.238**
Global Ratings - Interpersonal		
- Self	.201*	.226**
- Supervisor	.074	.036
- Peer	.042	.029
AF-Wide Ratings - Technical		
- Self	.214*	.184*
- Supervisor	.221*	.210*
- Peer	.228**	.217*
AF-Wide Ratings - Interpersonal		
- Self	.122	.108
- Supervisor	.060	.019
- Peer	.211*	.176*
ASVAB Composites		
Mechanical	.171	.077
Administrative	.072	.065
General	.125	.173
Electronic	.103	.089
New AFQT	.097	.164
TAFMS (Months in Service)	.286**	.251**
Technical Training Grade	.124	.191

Note. Self, N = 191; Supervisor, N = 188; Peer, N = 516.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX B-7: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 324X0)

	Hands-on Total	Interview Total
Number Times Performed[a]	.320**	.164
Last Time Performed (Weeks)[a]	.253*	.109
Task Experience Ratings[a]	.319**	.194
Task Ratings[a]		
- Self	.329**	.130
- Supervisor	.285**	.152
- Peer	.275**	.257*
Dimensional Ratings		
- Self	.302**	.086
- Supervisor	.286**	.221*
- Peer	.354**	.312**
Global Ratings - Technical		
- Self	.323**	.076
- Supervisor	.144	.142
- Peer	.277**	.248**
Global Ratings - Interpersonal		
- Self	-.033	-.123
- Supervisor	-.049	.014
- Peer	.097	.210*
AF-Wide Ratings - Technical		
- Self	.311**	.031
- Supervisor	.300**	.326**
- Peer	.270**	.245*
AF-Wide Ratings - Interpersonal		
- Self	.005	-.075
- Supervisor	.021	.047
- Peer	.110	.166
ASVAB Composites		
Mechanical	.288**	.275**
Administrative	.174	.168
General	.299**	.277**
Electronic	.369**	.289**
New AFQT	.285**	.229*
TAFMS (Months in Service)	.424**	.267**
Technical Training Grade	.372**	.282**
Job Knowledge Test (% Correct)	.587**	.473**

Note. Self, N = 138; Supervisor, N = 138; Peer, N = 331.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX B-8: CORRELATIONS BETWEEN JPMS VARIABLES AND WTPT SCORES (AFS 328X0)

	Hands-on Total	Interview Total
Number Times Performed[a]	.244*	.358**
Last Time Performed (Weeks)[a]	.061	.130
Task Experience Ratings[a]	.108	.183
Task Ratings[a]		
- Self	.333**	.367**
- Supervisor	.303*	.240*
- Peer	.138	.024
Dimensional Ratings		
- Self	.347**	.371**
- Supervisor	.333**	.332**
- Peer	.166	.053
Global Ratings - Technical		
- Self	.362**	.249*
- Supervisor	.253*	.255*
- Peer	.226	.118
Global Ratings - Interpersonal		
- Self	-.060	.048
- Supervisor	.002	-.057
- Peer	.056	.009
AF-Wide Ratings - Technical		
- Self	.387**	.352**
- Supervisor	.348**	.317**
- Peer	.240	.252*
AF-Wide Ratings - Interpersonal		
- Self	-.048	.046
- Supervisor	.080	-.044
- Peer	.150	.060
ASVAB Composites		
Mechanical	.309*	.317*
Administrative	-.003	-.015
General	.223	.311*
Electronic	.216	.258*
New AFQT	.285*	.330**
TAFMS (Months in Service)	.382**	.395**
Technical Training Grade	.184	.123

Note. Self, N = 97; Supervisor, N = 98; Peer, N = 188.
[a]Mean of ratings across all WTPT tasks.
*$p < .01$.
**$p < .001$.

APPENDIX C: WTPT TASK STATISTICS

APPENDIX C-1: WTPT TASK STATISTICS (AFS 423X5)[a]

Task Number	Mean	SD	Median
Overlap Tasks			
I251	1.34	2.33	0.00
H251	2.81	3.56	0.63
I209	5.92	2.01	6.36
H209	6.50	2.15	7.27
I300	6.05	2.06	6.62
H300	7.58	1.58	7.75
I264	4.26	1.57	4.32
H264	5.19	1.62	5.24
I421	2.98	2.02	2.06
H421	3.14	1.87	2.06
Unique Tasks			
H215	6.01	3.54	7.41
H155	4.64	2.29	4.77
I275	6.62	2.10	7.13
H154	8.11	1.56	8.67
I322	2.25	1.82	2.14
I340	5.87	2.96	6.00
H503	5.51	2.15	4.71
I120	4.95	2.02	5.05
H238	7.39	2.17	7.84
H260	6.39	3.68	8.37
I488	6.50	1.71	6.87
I477	7.03	1.65	7.27
I255	2.91	2.06	2.83
H179	3.06	1.91	3.12
I555	5.47	1.86	5.45
I286	5.37	2.29	5.59
H162	7.44	1.67	7.63
I181	2.74	2.66	2.12
H284	8.46	1.75	9.00
H446	2.86	2.65	2.50
H549	6.61	2.18	6.93

Note. Scores may exceed 10.00 due to equating process. Within each specialty, the scores were equated based on Phase I scoring to account for possible differences in difficulty across the Phase II/Phase III components. In this manner, each Phase II/Phase III portion had the same mean and variance and direct comparisons could be made within a specialty across all incumbents. Phase I refers to the portion of the WTPT that is applied to the entire specialty. Phase II tasks are specific to duty areas. All WTPTs had at least two phases with the exception of AFS 423X5 which required a single-phase approach. Phase III tasks are incumbent-unique based on location of work site (e.g., shop vs. flight line) and were included only in the AFS 426X2 WTPT.

[a]N = 261.

APPENDIX C-2: WTPT TASK STATISTICS (AFS 426X2)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I134	6.15	1.70	6.25
H134	7.21	2.09	7.44
I347	6.67	1.91	6.83
H347	7.02	1.72	7.15
I373	6.26	1.76	6.48
H373	8.48	1.90	8.18
Unique Tasks			
H301	6.22	1.92	6.20
H302	7.96	1.72	8.17
Phase II			
J-79[b]			
Overlap Tasks			
I353	6.94	2.91	7.55
H353	8.36	1.43	8.75
I360	5.95	2.18	5.80
H360	7.34	1.78	7.85
Unique Tasks			
I351	6.67	2.21	6.99
H363	8.60	1.30	9.08
I387	8.24	1.56	8.38
J-57[c]			
Overlap Tasks			
I353	5.89	1.76	5.87
H353	7.54	1.60	7.80
I360	7.07	1.90	7.64
H360	8.06	1.47	8.67
Unique Tasks			
I359	6.58	1.84	6.92
H363	7.83	1.25	8.03
I396	6.26	1.94	6.63
TF-33[d]			
Overlap Tasks			
I353	6.79	1.95	6.28
H353	6.99	1.54	6.71
I360	4.72	2.11	5.20
H360	6.09	2.30	6.30
Unique Tasks			
I359	6.17	2.47	6.59
H363	9.51	2.13	10.90
I387	7.77	2.08	8.22

Task Number	Mean	SD	Median
PHASE III			
J-79 Shop[e]			
H347	5.81	2.40	5.37
I247	3.18	2.73	2.89
I238	5.71	2.50	6.37
I239	3.57	3.12	2.19
H385	7.39	1.91	7.95
J-79 Flightline[f]			
H349	5.06	3.04	4.93
I319	5.00	1.97	4.85
I325	6.03	1.09	6.04
I328	7.26	2.66	7.36
H385	6.42	3.17	7.65
J-57 Shop[g]			
H349	6.49	1.91	6.31
I247	5.43	2.18	5.37
I238	6.56	1.95	6.74
I208	7.08	1.81	7.29
H346	6.77	2.20	6.79
J-57 Flightline[h]			
H349	6.23	1.99	6.44
I319	4.90	1.79	4.89
I325	4.97	1.61	5.24
I328	7.12	1.84	7.19
H171	7.30	1.68	7.74
TF-33 Shop[i]			
H349	4.24	1.95	4.57
I247	5.49	2.15	5.54
I238	6.53	2.27	6.98
I208	6.97	1.79	7.23
H346	8.21	2.33	7.58
TF-33 Flightline[j]			
H349	5.38	1.47	5.08
I319	4.24	2.17	3.92
I325	5.59	2.13	5.76
I328	6.85	3.17	7.89
H171	6.24	3.70	6.08

Note. Scores may exceed 10.00 due to equating process.
[a] $N = 255$.
[b] $N = 82$.
[c] $N = 89$.
[d] $N = 84$.
[e] $N = 49$.
[f] $N = 33$.
[g] $N = 32$.
[h] $N = 47$.
[i] $N = 46$.
[j] $N = 38$.

APPENDIX C-3: WTPT TASK STATISTICS (AFS 492X1)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I218	7.96	3.36	9.20
H218	7.72	3.55	8.98
I209	5.25	2.92	5.46
H209	8.17	3.26	8.96
I201	9.05	2.88	9.27
H201	9.62	2.42	9.26
I142	8.81	2.39	9.02
H142	8.85	2.34	8.95
Unique Tasks			
I192	7.42	2.39	7.83
H129	7.88	4.71	9.14
H143	7.69	2.59	7.85
I250	6.51	2.26	6.61
H258	8.49	2.28	8.30
H126	6.59	2.74	6.99
I216	9.02	2.47	9.23
IXXX	5.70	2.85	5.89
Phase II			
CISG[b]			
Overlap Tasks			
I173	3.83	3.36	3.91
H173	3.78	3.53	3.78
I184	7.94	2.84	7.93
H184	8.49	3.72	9.47
Unique Tasks			
H186	4.00	3.90	3.75
I169	7.82	4.22	8.80
H280	5.91	3.64	6.12
H237	6.44	2.63	6.38
I182	8.17	3.72	8.96
HWWW	5.09	4.49	4.20
GCCS[c]			
Overlap Tasks			
I166	7.93	1.87	8.57
H166	8.40	1.10	8.89
I197	7.96	2.03	8.61
H197	8.11	1.97	8.94
Unique Tasks			
H164	8.26	1.60	8.90
H183	7.01	2.11	8.08
H193	8.32	1.45	8.93
H175	7.71	1.72	8.23
I170	7.85	1.95	8.58
I236	8.05	1.61	8.56

Task Number	Mean	SD	Median
Giant Talk[d]			
Overlap Tasks			
I183	8.23	1.76	8.86
H183	8.21	1.32	8.76
I197	8.08	2.40	9.23
H197	8.33	1.87	8.95
Unique Tasks			
H164	7.91	1.96	8.89
HZZZ	8.50	.97	8.81
H193	8.34	1.74	8.92
H175	7.45	2.14	7.85
I170	6.50	2.89	7.38
I217	7.88	1.76	8.09

Note. Scores may exceed 10.00 due to equating process.
[a]N = 157.
[b]N = 68.
[c]N = 48.
[d]N = 41.

APPENDIX C-4: WTPT TASK STATISTICS (AFS 732X0)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I35	7.12	3.09	7.94
H35	7.38	3.13	8.56
I719	9.18	1.54	9.43
H719	9.17	1.57	9.55
I140	7.38	3.65	9.30
H140	7.24	3.73	9.39
Unique Tasks			
H131	8.29	2.13	8.47
H876	5.71	3.84	6.04
I121	9.94	1.21	9.82
I733	9.33	1.73	9.69
H116	7.73	2.53	7.74
Phase II			
C & T[b]			
Overlap Tasks			
I145	8.81	2.10	9.29
H145	7.92	1.85	8.13
I293	5.42	4.54	5.99
H293	6.18	4.58	8.47
I334	4.27	4.57	3.29
H334	4.80	4.71	5.71
Unique Tasks			
I335	6.96	3.40	7.37
I343	5.11	3.77	4.86
H157	8.72	2.54	9.45
I324	7.91	4.31	9.73
I296	5.71	4.13	6.63
HADD	5.26	4.37	5.55
H303	4.71	4.90	3.04
Manning[c]			
Overlap Tasks			
I476	7.87	3.21	9.58
H476	8.09	3.34	9.67
I1	8.04	2.32	9.58
H1	8.15	2.29	9.49
I466	8.08	2.71	9.67
H466	8.76	2.08	9.63
Unique Tasks			
I475	8.09	2.11	8.75
H436	1.62	2.86	0.45
I441	7.22	2.24	7.94
I437	4.95	2.62	4.75
I447	9.07	1.05	9.69
H472	9.13	1.68	9.62
H440	7.63	2.98	8.30

Task Number	Mean	SD	Median
Outbound[d]			
Overlap Tasks			
I415	8.22	1.86	9.20
H415	8.34	2.20	9.45
I396	8.56	1.60	9.14
H396	8.80	1.67	9.49
I380	6.01	3.02	6.20
H380	5.94	3.37	6.30
Unique Tasks			
I370	8.59	1.31	9.24
H398	8.04	1.72	8.44
H433	7.91	2.49	8.53
H357	7.84	2.49	8.30
H388	8.12	2.09	9.44
H389	7.87	2.94	9.06
H406	6.63	3.28	7.99
Separations[e]			
Overlap Tasks			
I835	9.08	1.10	9.34
H835	8.92	1.04	8.87
I874	5.76	3.71	7.67
H874	5.48	3.62	7.02
I876	5.00	4.03	4.07
H876	4.87	3.92	3.84
Unique Tasks			
I861	9.99	1.31	10.51
I852	5.32	3.16	5.32
I839	8.98	2.30	10.64
H884	9.27	.76	9.39
I840	7.88	2.74	8.61
H878	9.19	1.24	9.62
H889	4.36	3.60	5.38

Task Number	Mean	SD	Median
Records[f]			
Overlap Tasks			
I728	8.36	2.71	9.66
H728	8.59	2.77	9.95
Unique Tasks			
I734	8.04	1.63	8.24
I722	9.37	1.89	9.85
H739	7.09	3.48	8.74
H720	9.78	.83	10.02
I710	5.81	2.70	6.02
I711	8.30	2.02	9.38
I735	7.54	2.59	7.33
H713	8.52	2.02	9.29
H718	3.89	4.35	0.00

Note. Scores may exceed 10.00 due to equating process.
[a]N = 197.
[b]N = 31.
[c]N = 38.
[d]N = 46.
[e]N = 35.
[f]N = 47.

APPENDIX C-5: WTPT TASK STATISTICS (AFS 122XO)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I295	3.40	2.43	3.06
H295	6.07	2.86	6.74
I315	6.74	2.22	7.58
H315	7.60	1.91	7.60
I320	7.86	2.65	8.94
H320	9.48	1.83	9.96
I330	6.10	2.25	6.25
H330	6.92	2.36	7.37
I383	5.07	3.03	5.00
H383	6.91	3.33	8.35
I389	6.42	2.34	6.61
H389	8.11	2.10	8.59
Unique Tasks			
H199	7.96	2.98	9.16
H380	5.49	3.08	5.88
Phase II			
SAC[b]			
H303	6.67	2.09	6.73
H310	7.93	2.23	8.55
H349	7.24	2.05	7.70
H359	6.01	3.19	6.85
H398	6.69	2.35	7.14
MAC[c]			
H210	5.47	1.90	5.47
H224	6.33	2.55	5.63
H346	6.13	2.46	6.70
H379	6.14	2.51	6.39
H382	7.24	2.18	7.57
TAC[d]			
H278	8.93	1.69	9.15
H303	8.11	2.02	8.79
H311	6.18	4.10	8.24
H349	7.96	2.49	9.33
H483	4.20	4.63	0.00

Note. Scores may exceed 10.00 due to equating process.
[a] N = 195.
[b] N = 64.
[c] N = 82.
[d] N = 49.

APPENDIX C-6: WTPT TASK STATISTICS (AFS 272X0)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I172	6.29	2.03	6.46
H172	5.87	1.86	6.00
I274	7.63	2.56	8.53
H274	7.86	2.28	8.56
I293	8.88	2.05	9.90
H293	8.12	2.06	7.96
Unique Tasks			
I253	4.52	1.73	4.57
I260	7.07	1.61	7.07
H366	8.34	1.39	8.73
I305	3.50	2.43	3.69
I320	7.78	2.20	8.37
H278	7.30	1.49	7.45
H318	5.23	2.53	4.06
H319	5.67	2.45	5.74
Phase II Radar[b]			
Overlap Tasks			
I232	5.13	2.60	5.56
H232	5.16	2.64	5.52
I271	4.67	2.52	4.49
H271	5.06	2.52	5.59
I369	7.28	3.10	8.54
H369	7.84	2.87	9.17
Unique Tasks			
I277	6.15	2.07	6.06
I341	8.03	1.78	8.20
H364	7.76	2.63	7.93
I373	7.61	1.84	7.95
H339	8.85	1.49	9.47

Task Number	Mean	SD	Median
Tower[c]			
Overlap Tasks			
I406	8.06	1.94	8.77
H406	8.83	1.61	9.49
I405	7.27	2.52	8.00
H405	7.24	2.64	8.11
I381	4.72	1.94	4.15
H381	4.73	2.06	4.24
Unique Tasks			
I330	7.91	1.97	8.05
H389	8.04	1.85	8.17
H270	6.27	2.59	6.62
I398	7.92	1.54	8.05
H395	8.80	1.18	8.67

Note. Scores may exceed 10.00 due to equating process.
[a]N = 191.
[b]N = 139.
[c]N = 52.

APPENDIX C-7: WTPT TASK STATISTICS (AFS 324X0)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I211	8.88	.95	9.33
H211	9.38	1.21	9.69
I214	7.52	1.34	7.41
H214	8.17	1.31	8.37
I239	7.77	3.38	10.00
H239	9.17	1.79	9.66
I403	8.64	1.79	9.71
H403	9.73	1.29	9.76
I336	5.77	2.15	5.48
H336	7.27	2.08	7.52
Unique Tasks			
H268	8.13	2.14	8.35
I237	7.41	1.44	7.20
H270	6.69	3.39	7.12
H238	7.84	3.79	9.65
H436	6.10	2.44	6.24
H434	7.57	3.25	9.63
H764	9.67	.91	9.71
H373	9.45	1.36	9.71
H737	9.42	1.76	9.72
H781	3.41	2.25	3.12
H699	6.74	2.53	7.05
H645	8.75	2.34	9.67
Phase II			
K1/K2[b]			
H423	6.99	2.75	6.78
H391	8.71	2.71	9.93
H454	6.20	3.41	3.49
H452	5.15	2.16	4.95
H251	4.09	3.24	3.40
K3[c]			
H641	7.51	2.46	8.88
H672	7.66	2.68	9.05
H669	6.54	2.91	7.45
H679	7.79	2.02	7.99
H748	8.98	1.86	9.66
K4[d]			
H827	8.71	3.05	10.05
H845	3.05	.75	3.12
H630	8.43	2.99	9.45
H846	8.04	1.87	7.75
H260	6.23	5.34	7.05

Task Number	Mean	SD	Median
K5/K6[e]			
I971	7.01	1.98	7.02
H971	8.62	1.32	9.11
I934	3.79	3.26	4.00
I1006	6.87	2.65	7.87
H1006	7.79	1.80	8.16
I1021	8.27	2.32	8.73
H1021	8.74	1.80	9.54
K8[f]			
H506	8.00	3.29	9.30
H509	6.77	3.93	9.09
H499	7.38	4.06	9.98
H561	4.34	4.92	0.00
H560	1.69	3.00	0.00

Note. Scores may exceed 10.00 due to equating process.
[a]N = 138.
[b]N = 40.
[c]N = 51.
[d]N = 4.
[e]N = 19.
[f]N = 24.

APPENDIX C-8: WTPT TASK STATISTICS (AFS 328X0)

Task Number	Mean	SD	Median
Phase I[a]			
Overlap Tasks			
I160	8.56	1.58	8.42
H160	9.12	1.36	8.98
I232	6.48	1.89	6.54
H232	7.98	1.99	8.46
I233	8.15	1.65	8.27
H233	9.31	1.33	9.24
I234	5.58	2.71	5.52
H234	6.42	3.00	6.58
Unique Tasks			
I163	7.28	2.16	7.36
H173	5.95	3.10	6.36
I218	5.37	3.07	5.40
H238	6.07	2.48	5.70
I240	6.63	2.90	7.62
I244	9.16	1.54	9.16
H250	8.96	2.18	9.22
I253	8.73	3.48	9.56
I257	7.66	2.63	8.10
H257	8.92	2.32	9.29
H260	8.11	1.76	8.55
I539	5.28	1.81	5.28
Phase II			
SAC[b]			
Overlap Tasks			
I458	4.98	3.23	6.14
H458	6.87	3.17	9.01
Unique Tasks			
H439	8.76	0.51	8.92
H448	6.76	3.61	8.89
H459	7.70	2.68	8.89
I536	7.41	1.46	7.51
I540	5.95	3.70	8.27
MAC[c]			
Overlap Tasks			
I459	5.84	3.89	5.74
H459	9.65	1.19	9.88
Unique Tasks			
H325	5.83	3.95	7.08
H327	9.10	1.50	9.50
I434	8.66	1.24	8.65
H445	9.43	1.32	9.60
I451	2.67	2.50	1.80

Task Number	Mean	SD	Median
TAC[d]			
Overlap Tasks			
I258	4.57	4.30	4.45
H258	4.42	4.45	3.99
Unique Tasks			
I226	3.79	2.44	3.95
H229	8.94	2.73	9.28
H231	9.98	2.46	11.17
I245	7.74	1.95	8.28
H249	9.69	3.30	10.97

Note. Scores may exceed 10.00 due to equating process.
[a]N = 98.
[b]N = 35.
[c]N = 32.
[d]N = 31.